Atomic Aggregation on 3D Gaussian Splatting
Fan Chen, Xin Peng, Keyi Zhang
Master of Science in Applied Computing, University of Toronto

Motivation
3D Gaussian Splatting\(^1\) achieves real-time rendering by leveraging GPU rasterization pipeline. However, the training of these models remains a computationally demanding process. By an in-depth performance analysis of 3DGS using Nvidia profiler\(^2\), we identified the **atomic updates as a significant bottleneck** at the gradient computation step in backward pass.

Goal: Enhance the training efficiency by consolidating atomic instructions within the backward pass.

Key Observations:
1. Threads within the same warp exclusively update identical memory locations.
2. Atomic updates are solely performed by a subset of threads within a warp.

Key Contributions
- Conducted an exhaustive performance analysis on the training pipeline of 3DGS and discern atomic updates as a pivotal bottleneck.
- Introduced a software approach that uses warp-level reduction to reduce the number of atomic updates.
- Evaluated our approach on 3DGS application and demonstrate significant speed up.

Related Work
**Atomic Processing in GPU:**
- Individual threads perform atomic updates on specific data or memory locations to maintain data integrity.
- Ensures correct and conflict-free modifications in a parallel computing environment where multiple threads may simultaneously access the same memory location.

**Warp Reduction:**
- Using the __shift_down_sync primitive enables fast and direct data exchange between thread registers, which is more efficient than using shared memory\(^3\). This method can be used to accumulate results in a specific thread.

```
#define FULL_MASK offset
2
for (int offset = 16; offset > 0; offset /= 2)
val = __shll_down_sync(FULL_MASK, val, offset)
```

- The breakdown of the number of cycles a warp is stalled per instruction on the NVIDIA RTX 4090 and 3060 GPUs. Left-top is the result of baseline, left-bottom is the result of our approach.
- Below is the load store unit utilization for both the baseline and our proposed approach.

References